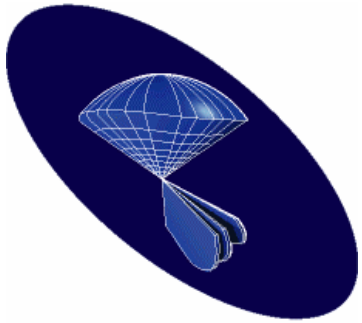
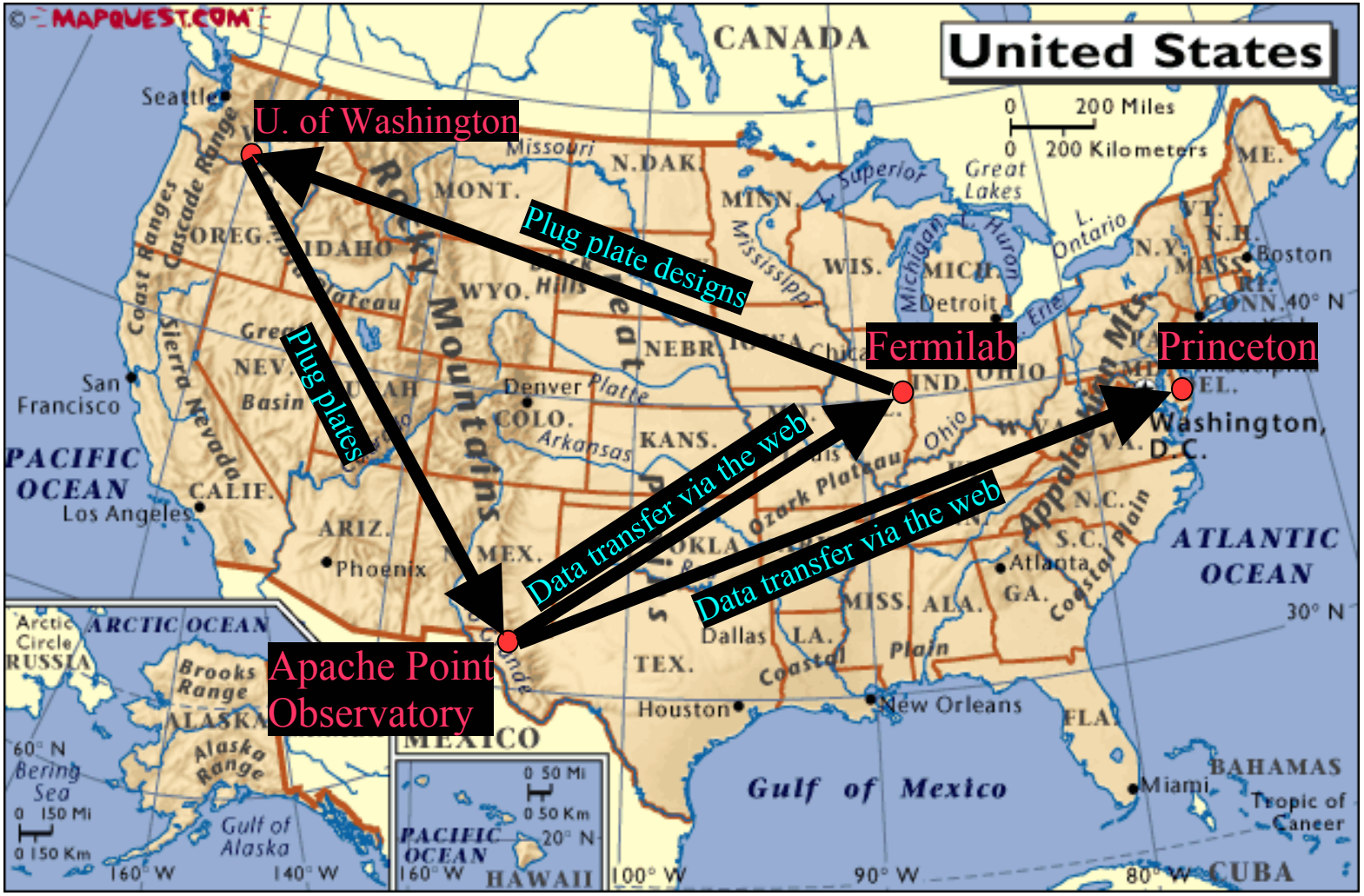


Data Acquisition & Distribution; and Archiving Plans

Bill Boroski

SDSS-II First Year Review
National Science Foundation
August 7, 2006

Data Acquisition & Processing Operations



Data Acquisition / Observatory Operations

- Observing operations at APO
  - Performance metrics remain solid

    Ave. imaging efficiency = 87% (Aug-Jan); 72% (Mar-Jun)

    Ave. spectro efficiency = 67%

    Ave. system uptime = 96%

  - Data transfers via 15 Mbps microwave link
    - Only two disruptions in the first 9 months of operation
    - Recovery was relatively quick; no data lost

  - SN compute cluster for on-the-mountain reductions

  - Fire protection upgrades

Data Processing Operations

- Data Processing Operations at Fermilab
  - Production processing of all Legacy, SEGUE and Supernova Survey data; target selection and plate design
  - Operations continue to run smoothly and efficiently
  - Implemented automated method for pulling data from APO and verifying data integrity
  - Implemented scripts to “crawl” entire spinning data set to detect file corruption, missing files, etc.

- Data Processing Operations at Princeton
  - DP hardware cluster installed and commissioned
  - Imaging and spectro data are being reduced
  - Imaging reductions being used to support Photo pipeline and photometric calibration work
  - Spectro reductions being used in Spectro v5 pipeline development

Supernova Survey Data Release 1

- First public release of SN data occurred in January 2006
  - Contains data from 72 imaging runs obtained during the fall 2005 observing campaign (Sep 1 through Nov 30, 2005)
  - Contains an additional 16 imaging runs obtained during the fall 2004 season
  - Data taken on a 2.5 degree wide region along the celestial equator, from -60 < RA < 60 (SDSS Stripe 82)

- URL (linked off the SDSS home page): http://www.sdss.org/drsn1/DRSN1_data_release.html

- Web page provides access to corrected frames and uncalibrated catalogs via pared-down DAS script
  - Also provides list of available runs with information regarding data quality

- Future data releases will occur incrementally during each fall SN observing season
  - Releases will occur roughly 4-6 weeks after data collection

Data Release 5

- Contains survey-quality Legacy and SEGUE data collected through June 30, 2005.

Imaging	
Footprint Area	8,000 sq. deg.
Imaging Catalog	215 million objects
Data Volume	
Images	9.0 TB
Catalogs (DAS, fits format) (CAS, SQL DB)	 1.8 TB 3.6 TB

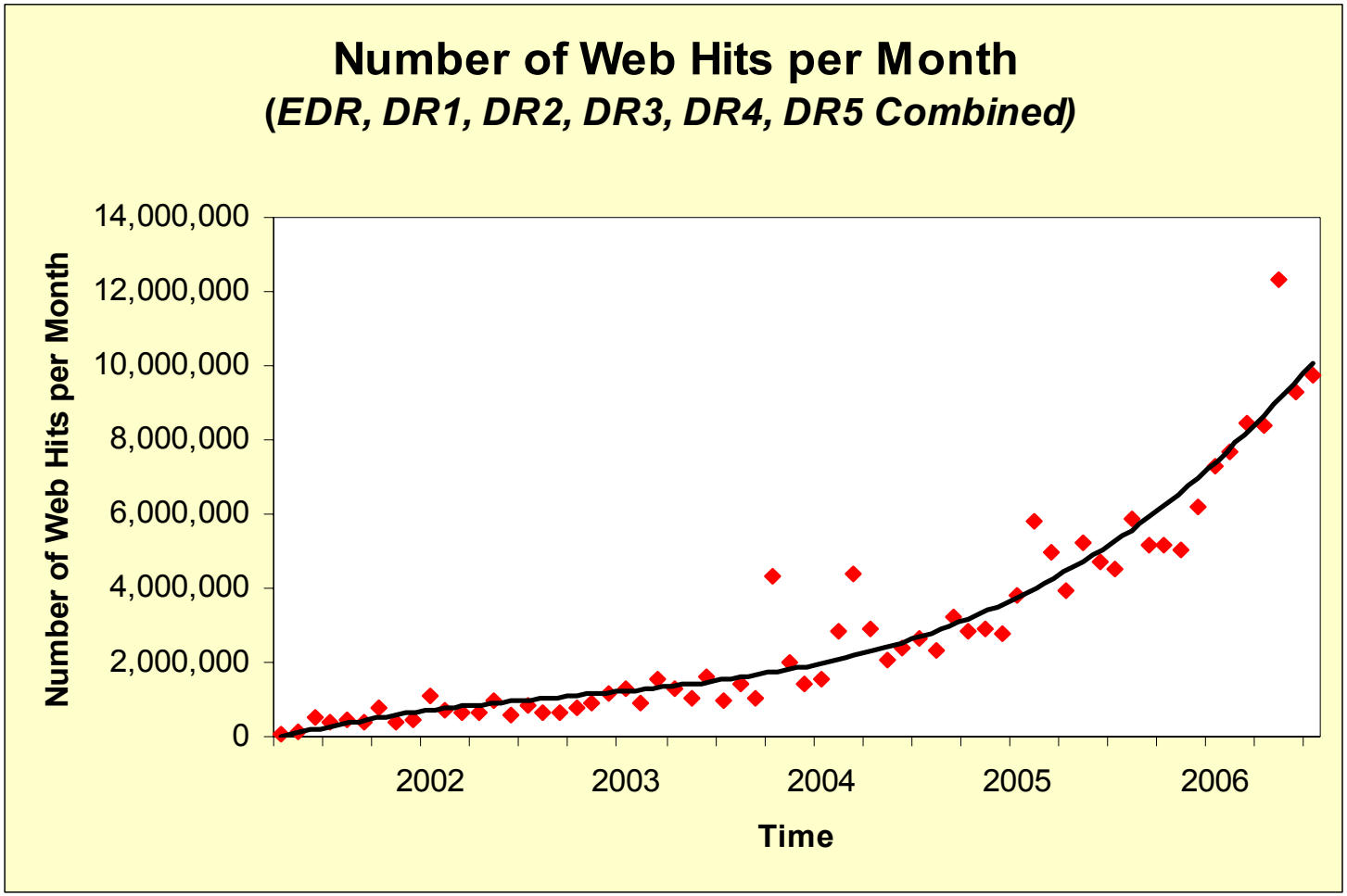
Spectroscopy	
Spectro Area	5,740 sq. deg.
Total # of spectra	1,048,960
Galaxies	674,749
Quasars (redshift < 2.3)	79,934
Quasars (redshift > 2.3)	11,217
Stars	154,925
M stars and later	60,808
Sky spectra	55,555
Unknown	12,312

- Also contains data from 361 “extra” and “special” plates
  - *Repeat observations, observations of spectro targets selected by collaboration members for specialized science programs.*
- The “final” release of the SDSS Survey. All data now in public domain.

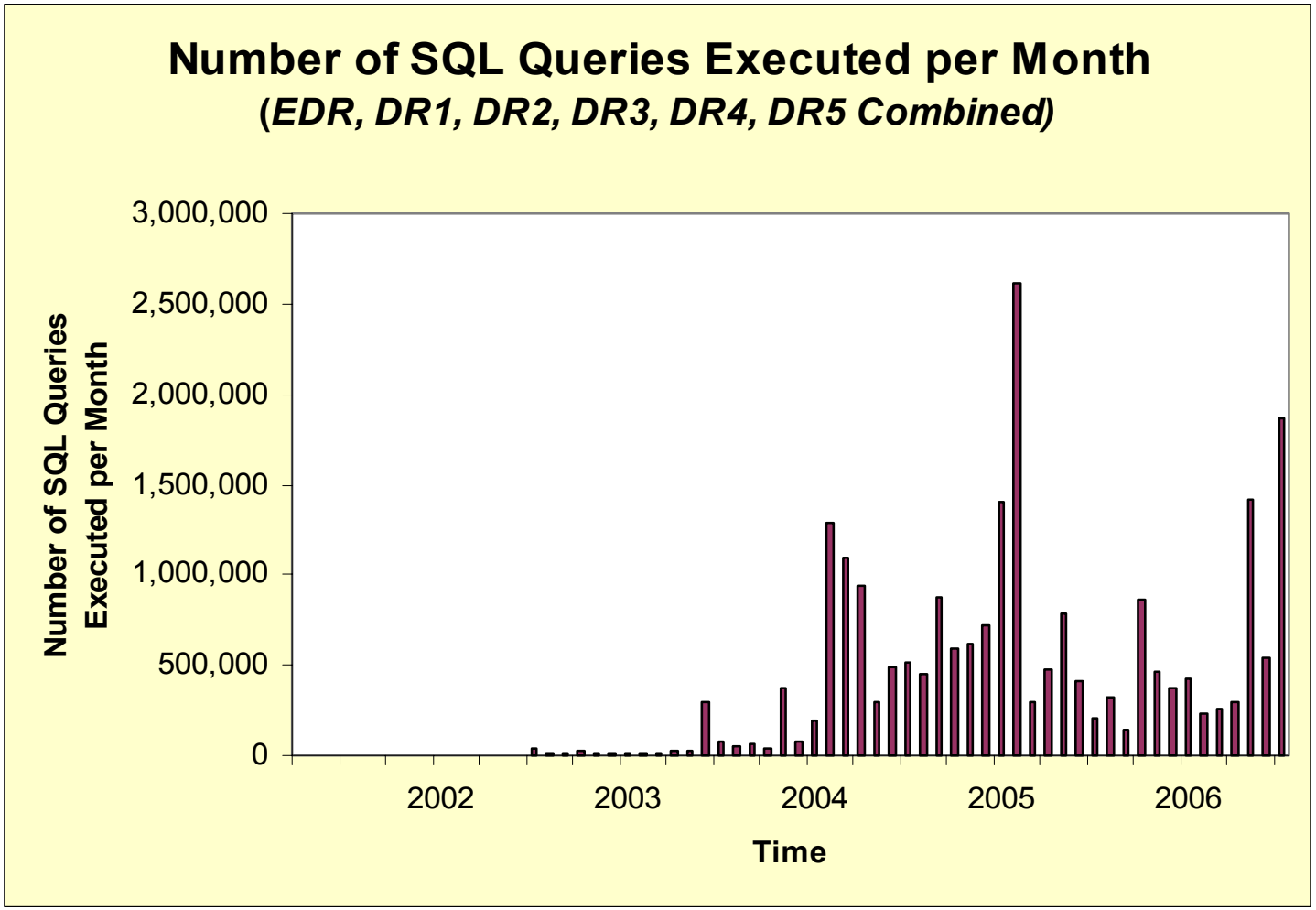
Data Release 5 (cont'd)

- October 21, 2005 - *initial DR5 CAS and DAS released to collaboration*
- May 4, 2006 – *DAS and enhanced CAS released to collaboration*
  - Photometric redshifts of galaxies
    - computed by two different groups within the collaboration
  - Detailed coverage masks
    - allow LSS researchers to easily calculate power spectrum and related quantities;
    - Allow computation of the area covered by statistical samples, such as the quasar sample in DR5.
  - QSO catalog tables
    - Master list of everything that "smells" like a quasar, with vital signs gathered from different data sources (i.e., FIRST, ROSAT, Stetson, USNO, and USNO-B catalogs).
  - runQA for target photometry (also bug fix)
  - Regenerated JPEGs to fix clipping error in some images.
- June 28, 2006: Enhanced version made available to the public
  - Usage varied between 80,000 and 100,000 SQL queries per day in the week immediately following the public release.

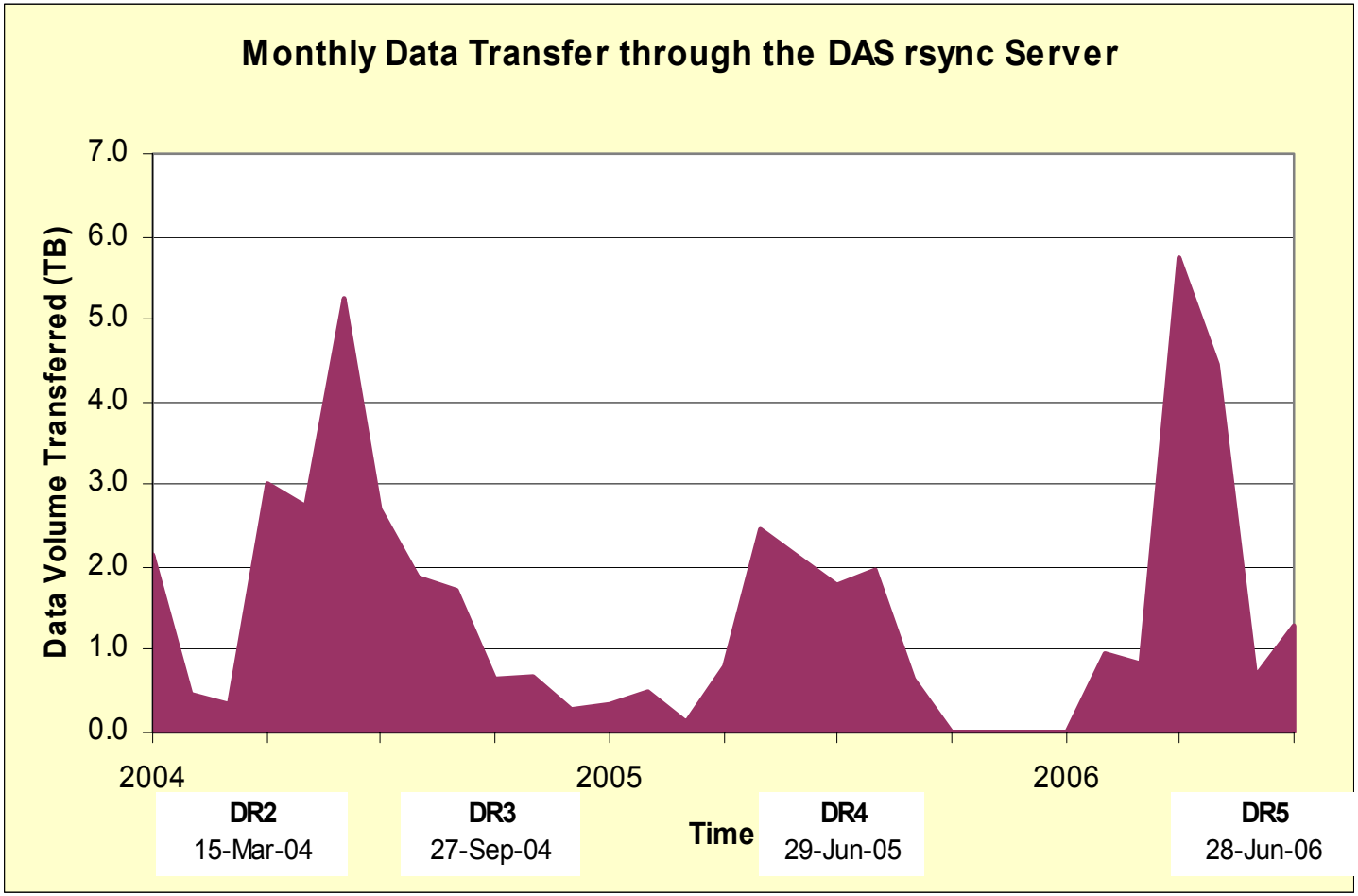
SkyServer Data Usage Statistics



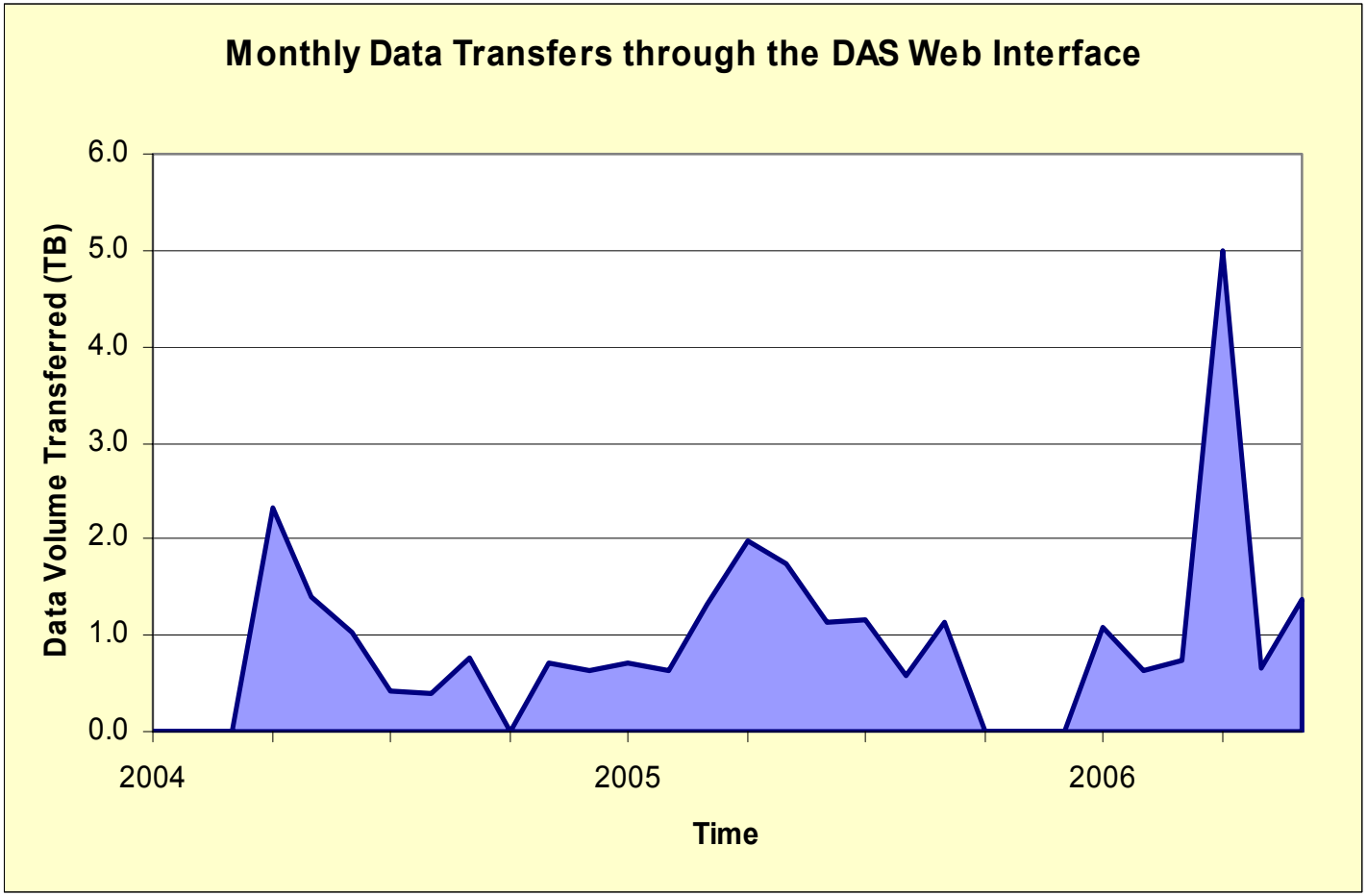
SkyServer Data Usage Statistics



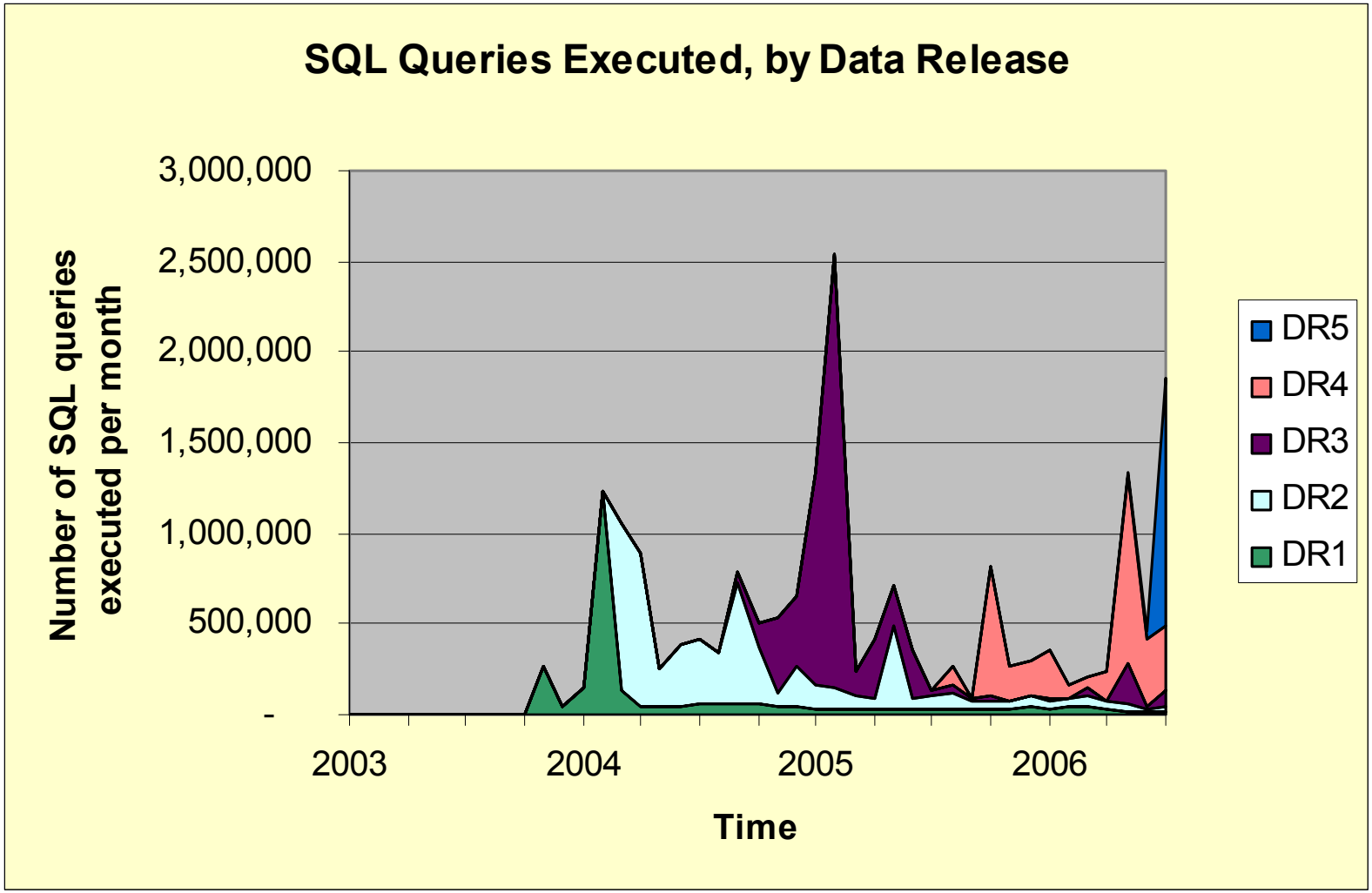
DAS Data Usage Statistics



DAS Data Usage Statistics



Data Usage by Release



Looking ahead to Data Release 6

- Public release scheduled for July 1, 2007
  - Legacy and SEGUE data collected through 30-Jun-2006
- DR6 will incorporate modest data model changes, such as:
  - Add table containing SEGUE stellar parameters
  - Add primary and secondary target fields to photoObjAll and specObjAll tables to accommodate SEGUE and “other” plates
  - Extract photometry for targets into new targPhotoObjAll table in BESTDR6, which allows us to eliminate TARG DB from some servers
  - Develop sqlFits2CSV code to read in Pan-STARRS object tables and build CSV files for DB loading
- Will load and deploy using MS SQL Server 2005
- Code modification and testing currently underway
- Data loading will commence in 4-6 weeks, with collab release occurring in late fall.

Runs Database

- CAS-style database that is being loaded with all imaging runs obtained over the course of the survey, regardless of data quality
- Updated version released to the collaboration on July 27, 2006
  - Contains 370 imaging runs
- Incremental releases are being made every 6-8 weeks
  - Each increment has contained of order 40-60 new runs
- Collaboration DAS access is also available for these runs
  - FITS format corrected frames
  - Uncalibrated (fpObjc*, asTrans*) and calibrated (tsObj*, tsField*) data.
  - Collab access via rsync and wget
- We intend to make the RunsDB available to the public at the time of the final SDSS-II data release, when all data will be in the public domain.

Interim Archive Stewardship

- A formal plan for the long-term stewardship of the SDSS data set is still being formulated

- The possibility of an interim arrangement has been discussed with senior Fermilab management
  - Lab management has expressed a willingness to serve as interim host until a more permanent steward is identified – timescale TBD;
  - Initial level of stewardship would include:
    - Maintaining data integrity
      - Retention of multiple copies; automated scanning to detect data corruption.
    - Maintaining existing systems (SkyServer, CAS, DAS, sdss.org, etc.)
      - Replace failed hardware, apply security patches, deploy OS upgrades, etc.
      - Helpdesk support?? (level to be determined)
  - Estimated resource requirements
    - One FTE (HW/DBA support)
    - $100K/yr for equipment, materials and supplies.

Long-term Stewardship

- Curation
  - *Mandatory*
    - Format conversion (e.g., OS upgrades, etc.)
    - Platform migration (Database, OS, etc.)
  - *Value-added*
    - Errata
    - Bug fixes
    - Annotations
      - First tier – team
      - Second tier – archiving centers
      - Third tier – community
    - Virtual Observatory (VO) compliance

Long-Term Stewardship
Preliminary Thoughts on Resource Requirements

- Operations Support
  - System maintenance
    - Hardware-level support
    - Application-level DBA support
    - System performance monitoring
  - Help desk
  - Logging
    - Usage statistics

- <u>Estimated Resource Requirements</u>
  - Number of boxes: 10-20
  - Level of effort:
    - 1 FTE for helpdesk support
    - 1 FTE for hardware/software support
    - 0.5 FTE for technical support (scientists, etc.)
  - Budget for ongoing maintenance and support costs
    - Disk replacements, software upgrades, server upgrades, etc.
    - ~$100K, dropping off in future years